Assignment #1 for Big Data Mining – COSC 526

0. Download code made available at https://github.com/ssrangan/gm-sparql

1. Two small datasets will be made available on the class website. One those datasets uses the code downloaded in Step 1 to compute (i) degree, (ii) PageRank (iii) triangle count (iv) node eccentricity for each node. (10 points)

2. Implement Random Walk algorithm using SPARQL using Python. Try to be as parallel-minded and SPARQL-friendly as possible during your implementation. (40 points)
   a. Show that the random walk algorithm can be used to navigate the “data maze” in an intelligent way and can help identify relevant/useful/meaningful paths.
   b. You can be creative about designing the measure of “usefulness/relevance and meaningfulness by using graph metrics for the graph transition probability, but if you do not know where to start – try using the graph metrics computed in Question 1.

3. Modify the algorithm in Step 2 to do ‘k’-random walks at a time in parallel (40 points)

4. Is implementation for Step 3 faster than the serial walker? Explain Why. (10 points)

5. Design “your own” method to rank the parallel walks for “interestingness”/”novelty” and qualitatively compare your results with results from Step 2. (BONUS: 10 points)

Submission Instructions

Assignment submission:
Report.pdf or Report.doc

1. Be brief but powerful. A good answer is typically between 100 -200 words per Question.

Code submission:

2. Your code should be similar in flavor to the EAGLE toolbox – one core file that implements the algorithm and one wrapper (unit test file) that calls the implementation. Full credit for the assignment only if the code executes.
3. Place code for each question in a separate folder q2, q3, q5, etc.
4. Zip all the folders with the sample output results in a text file and submit.